Large Language Models can Accurately Predict Searcher Preferences

Year: 2023 **|** *Citations: 213* **|** *Authors: Paul Thomas, S. Spielman, Nick Craswell*

Abstract

Much of the evaluation and tuning of a search system relies on relevance labels---annotations that say whether a document is useful for a given search and searcher. Ideally these come from real searchers, but it is hard to collect this data at scale, so typical experiments rely on third-party labellers who may or may not produce accurate annotations. Label quality is managed with ongoing auditing, training, and monitoring. We discuss an alternative approach. We take careful feedback from real searchers and use this to select a large language model (LLM), and prompt, that agrees with this feedback; the LLM can then produce labels at scale. Our experiments show LLMs are as accurate as human labellers and as useful for finding the best systems and hardest queries. LLM performance varies with prompt features, but also varies unpredictably with simple paraphrases. This unpredictability reinforces the need for high-quality "gold" labels.